IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 14, 2019, accepted January 29, 2019. Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2019.2898353*

# Lexical Dependent Emotion Detection Using Synthetic Speech Reference

**REZA LOTFIAN, (Student Member, IEEE), AND CARLOS BUSSO, (Senior Member, IEEE)**
Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX 75080, USA

Corresponding author: Carlos Busso (busso@utdallas.edu)

This work was supported in part by the National Science Foundation (NSF) under Grant IIS-1329659 and CAREER Grant IIS-1453781.

**ABSTRACT** This paper aims to create neutral reference models from synthetic speech to contrast the emotional content of a speech signal. Modeling emotional behaviors is a challenging task due to the variability in perceiving and describing emotions. Previous studies have indicated that relative assessments are more reliable than absolute assessments. These studies suggest that having a reference signal with known emotional content (e.g., neutral emotion) to compare a target sentence may produce more reliable metrics to identify emotional segments. Ideally, we would like to have an emotionally neutral sentence with the same lexical content as the target sentence where their contents are timely aligned. In this fictitious scenario, we would be able to identify localized emotional cues by contrasting frame-by-frame the acoustic features of the target and reference sentences. This paper explores the idea of building these reference sentences leveraging the advances in speech synthesis. This paper builds a synthetic speech signal that conveys the same lexical information and is timely aligned with the target sentence in the database. Since it is expected that a single synthetic speech will not capture the full range of variability observed in neutral speech, we build multiple synthetic sentences using various voices and text-to-speech approaches. This paper analyzes whether the synthesized signals provide valid template references to describe neutral speech using feature analysis and perceptual evaluation. Finally, we demonstrate how this framework can be used in emotion recognition, achieving improvements over classifiers trained with the state-of-the-art features in detecting low versus high levels of arousal and valence.

**INDEX TERMS** Emotional speech analysis, emotional speech recognition, synthesis of speech, feature normalization.

## I. INTRODUCTION

Emotion plays an important role in interpersonal human interaction [1]. *Human-machine interfaces* (HMIs) will benefit from incorporating emotional capabilities to recognize the affective states of users. Studying and understanding the emotional modulation conveyed on expressive speech is an important step toward designing robust machine learning frameworks that exploit the underlying production of emotional speech. Emotional speech presents localized cues that a robust system should consider [2]–[5]. This paper proposes a novel method based on reference models built with synthetic speech to quantify deviations from neutral speech.

Quantifying emotional cues conveyed in speech is a challenging problem, not only for machines [6]–[8] but also for human [9]. The challenges arise due to differences in emotion perception and ambiguous descriptors to properly represent the emotional behaviors [10], [11]. While assigning absolute emotional attributes commonly leads to disagreements, we are more reliable in comparing the emotional content between stimuli (e.g., sentence one is happier than sentence two) [12], [13]. These observations have motivated the development of preference learning in affective computing, where the task is to rank emotions according to predefined scales [14]–[19]. An interesting alternative, motivated by these studies, is to have a reference sentence with a known emotional profile that is used to contrast the target sentence. If the reference sentence is emotionally neutral, in particular, the comparison can serve as an effective framework to

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

quantify deviations from neutral patterns regardless of the actual emotion conveyed on the target sentence.

The ideal scenario for this framework is when the reference sentence conveys exactly the same lexical information as the target sentence, and their contents are timely synchronized. In this fictitious scenario, we can directly compare frame-by-frame the acoustic properties of both signals, highlighting emotionally salient segments that deviate from neutral speech. Can advances in *text-to-speech* (TTS) systems provide a systematic framework to build these neural reference sentences? Our preliminary analysis showed the feasibility of this idea [20]. This study further explores this question, proposing a novel approach to build a robust emotion recognition system that exploits the underlying nonuniform externalization process of expressive behaviors. We build a synthetic speech signal that conveys the same lexical information and is timely aligned with the target sentence in the database. The approach consists of using the spoken message conveyed in the sentence to synthesize a reference signal. The phonetic transcriptions of the synthesized signal and the target sentence are then aligned, generating a reference temporarily aligned with the original sentence. Since it is expected that a single synthetic speech will not capture the full range of variability observed in neutral speech, we produce different neutral synthetic realizations using various voices and TTS models (e.g., family of synthesized signals).

We explore the hypothesis that synthesized speech provides a valid template reference to describe the acoustic properties of neutral speech. The proposed approach consists of comparing the property of neutral, synthetic and emotional speech with feature analysis and perceptual evaluations. We use a database recorded to build *automatic speech recognition* (ASR) systems to represent the intrinsic variability observed on neutral speech. We consider the synthesized signal both before and after the temporal alignment to understand the distortions introduced by the alignment process. The analysis identifies the features from synthesized speech that better represent the acoustic properties of neutral speech. Likewise, we conduct perceptual evaluations to assess the emotional percepts of neutral, synthetic and emotional speech. The emotional subjective evaluations are compared between speech groups (neutral, synthetic and emotional speech). The subjective evaluation indicates that the synthetic speech and time-aligned synthetic speech are mainly perceived as neutral, confirming the assumption that these signals can be used as neutral references.

After creating the synthetic reference signals and validating their potential to represent neutral speech, we demonstrate one potential use in the area of speech emotion recognition. The synthesized speech references are used to contrast the localized emotional content of a target sentence by using a lexical normalization approach. The method is a modified version of the whitening transformation introduced in Mariooryad and Busso [21], where the synthetic reference signals are used to attenuate the lexical information on the original speech. By reducing the uncertainty introduced by the lexical content, we expect to increase the relevance between the normalized acoustic features and emotion. The classification evaluation shows improvements when we include features extracted from the normalized speech, demonstrating the merits of using synthesized speech references in speech emotion recognition.

The rest of the paper is organized as follows. Section II summarizes important contributions from previous studies in the context of the proposed framework and the databases used for the analysis. Section III describes our approach to generate synthetic reference sentences and how they are used to contrast the emotional content of a sentence. Section IV validates the use of synthetic speech to represent neutral speech with acoustic analysis and perceptual evaluations. Section V presents the experimental results of emotion classification demonstrating the effectiveness of the proposed framework. Section VII concludes the paper with discussion, future directions and final remarks.

## II. BACKGROUND AND RESOURCES

### A. RELATED WORK

Acoustic features have been largely used for emotion recognition [22]–[24]. The most common approach is to derive global statistics at the sentence level from prosodic and spectral features [25]. Some studies have proposed to recognize emotions using smaller units such as words or chunks, to capture emotional variability within a sentence [26]. Cowie *et al.* [27] stated that emotions either gradually or sharply shift over time. They even designed an annotation software, FEELTRACE, to continuously track the emotional variations within a sentence (see Sec. II-B for details about this toolkit). One important factor is that there are salient words that conveys more intense emotions [24]. In fact, Whissel [28] proposed the *dictionary of affect* to measure the emotional content of the words. Common words were labeled using the dimensions pleasantness, activation and imagery. Another factor is the presence of localized trends for specific emotions. For example, the pitch slope tends to increase at the end of happy sentences [29].

The nonuniform emotional modulation is also observed at the phoneme level. We have studied the phoneme level patterns for angry, happy, sad and neutral sentences [4]. The vowel triangle was estimated, which describes the first and second formant frequencies for the vowels /iy/, /uw/ and /aa/. The study showed that low vowels (e.g., /aa/), with less restricted tongue position present stronger emotional modulation than high vowels (e.g., /iy/). Similar observations were reported by Goudbeek *et al.* [30]. Likewise, we found clear emotional differences in the spectral properties observed across broad phonetic classes (e.g., frontal vowel, fricatives, diphthong and nasal sound, etc.) [5]. We observed higher emotional modulation during frontal vowels than during nasal phonemes. This result is explained by the limited flexibility in the speech production system to generate nasal sounds. Altogether, these results suggest that articulatory constrains limit the degree of freedom to convey emotions. Therefore, it is expected that some segments will present stronger emotional modulation. As an aside, we have observed that facial expressions have higher emotional modulation during the temporal segments in which the acoustic features are physically constrained [31]. This result indicates that emotions are also modulated across modalities.

Instead of creating models for individual phoneme classes, some studies have attempted to attenuate the lexical variability with feature normalization. Mariooryad and Busso [21] proposed a feature normalization technique based on the whitening transformation to accomplish this goal. For a given acoustic feature (e.g., F0 contour), their method builds a trajectory model for each phoneme, which is represented as a $N$ dimensional vector by interpolating and resampling the original shape of the feature. The trajectory model is used to perform a whitening transformation where their parameters are applied per phoneme. The study showed a 4.1% classification performance improvement by reducing the variability associated with the lexical content. A limitation of this approach is the discontinuities in the normalized features due to the separate transformation applied to each phoneme. This study aims to build reference models for the entire sentence to attenuate the lexical variability, avoiding the discontinuities between phonemes.

The primary contribution of this paper is to introduce the use of synthetic speech as a reference of neutral speech to build a model to contrast the emotional content of a target sentence. This is not the first time that synthetic speech has been used in emotion recognition. Schuller and Burkhardt [32] proposed to use emotional synthetic speech to address the problem of data sparseness in emotion recognition. Their group extended that work, showing the benefit of training and adapting acoustic models using synthesized speech along with human speech, especially for cross-corpus applications [33]. These studies are radically different from our work, since they used emotional TTS to increase the training database. Instead, our goals in using TTS are to:

- Create neutral synthetic reference signals that convey the same lexical information and are timely aligned with a target sentence
- Evaluate the hypothesis that synthesized speech provides a valid template reference to describe neutral speech
- Contrast the localized emotional content of a target sentence with the reference synthetic speech, improving classification performance

The proposed approach is very novel with important implications in affective computing beyond speech emotion recognition. The use of synthetic speech to contrast emotional cues is an elegant formulation for the analysis of emotions. Current approaches often deal with machine learning algorithms where the only criterion is classification performance. Very often, these models cannot be used to interpret the predictions. With the proposed approach, we create a family of synthetic speech signals, which is used as a reference to contrast expressive speech. We can directly evaluate the deviations at the segmental level between the expected acoustic features (synthetic speech), with the acoustic features of the target speech. This formulation can provide a better understanding of the externalization of emotion in speech.

### B. DATABASES

The study relies on two databases. The first corpus is the SEMAINE database [34]. This corpus contains annotated multimodal recordings of emotionally colored conversations between two parties, a *user* which is always a human, and an *operator*. The operator, which can be a virtual agent or a human, takes four personalities to induce emotional reactions on the user: Poppy who is happy, Spike who is angry, Prudence who is reasonable and Obadiah who is gloomy. This study only uses the Solid SAL subset, where the operators are humans portraying the given personalities. This set includes 95 sessions collected from 20 subjects, where each session is approximately five minute long. In total, we use 2,773 speaking turns.

The sessions are emotionally annotated using the FEELTRACE toolkit [27] by multiple evaluators. FEELTRACE records continuous traces describing the values of a given emotional attribute across time. The evaluator watches a video, judges the emotional content, and annotates his/her instantaneous reactions by moving the mouse's cursor over an appropriate area in the interface. The axes in the display represent the target attributes that evaluators are asked to annotate. To compensate for the reaction lag of the evaluators in annotating the emotional content (i.e., the delay of the evaluators in reacting to the emotional content in the sentence), we rely on the methodology proposed by Mariooryad and Busso [35], [36]. While the database provides annotations for several emotional attributes, this study only uses valence (negative versus positive) and arousal (calm versus active). The study considers segment-based analysis, where we estimate the average of the emotional traces for each speaking turn (i.e., average across the evaluators, and across the duration of the turn).

Unlike the SEMAINE database, the second corpus only contains emotionally neutral sentences serving as our neutral (i.e., non-emotional) reference database. We use this corpus to compare the naturalness of synthesized speech in our analysis (see Sec. IV). We rely on the Wall Street Journal-based Continuous Speech recognition Corpus Phase II database [37], which we refer to as WSJ. While the corpus has also read speech, we only uses the spontaneous set which comprises 8,104 sentences, uttered by 50 different journalists.

### C. FEATURE EXTRACTION

The proposed framework consists of contrasting acoustic features of synthetic reference signals and the target sentence. We can implement this framework with different acoustic features (e.g., prosodic, spectral, voice quality). For this purpose, we use the exhaustive feature set defined for the INTERSPEECH 2013 *computational paralinguistics challenge* (ComParE) [38], extracted with the OpenSMILE toolkit [39]. This feature set is defined by estimating *low level*

**TABLE 1.** The set of frame-level acoustic features in the ComParE feature set [38], referred to as *low level descriptors* (LLDs).

| **Spectral LLDs** |
|---|
| RASTA-style filtered auditory spectrum bands 1-26 (0-8kHz) |
| MFCCs 1-14 |
| Spectral energy 25-650Hz, 1k-4kHz |
| Spectral roll-off point 0.25, 0.50, 0.75, 0.90 |
| Spectral flux, entropy, variance, skewness, kurtosis, slope |
| Slope, Psychoacoustic Sharpness, Harmonicity |
| **Energy related LLDs** |
| Sum of auditory spectrum (loudness) |
| Sum of RASTA-style filtered auditory spectrum |
| RMS Energy |
| Zero-crossing rate |
| **Voice LLDs** |
| F0 |
| Probability of voicing |
| Logarithmic HNR |
| Jitter (local, delta) |
| Shimmer (local) |

**TABLE 2.** The set of sentence-level functionals in the ComParE feature set [38], extracted from the LLDs (see Table 1).

| **Base functionals applied to LLD and $\triangle$ LLD** |
|---|
| Quartiles 1-3 |
| 3 inter-quartile ranges |
| 1% percentile (≈min), 99% percentile (≈max) |
| Position of min/max |
| Percentile range 1%-99% |
| Arithmetic mean, Root Quadratic Mean |
| Standard deviation,Skewness, kurtosis |
| Contour centroid, Flatness |
| Relative duration signal is above/below 25/ 50/ 75/ 90% range |
| Relative duration signal is rising/falling |
| Relative duration LLD has positive/negative curvature |
| Gain of linear prediction (LP) |
| LP coefficients 1-5 |
| **Base functionals applied to LLD only** |
| Mean of peak distances |
| Standard deviation of peak distances |
| Mean value of peaks |
| Mean value of peaks-arithmetic mean |
| Mean / Standard Deviation of rising/falling Slopes |
| Mean/ Standard Deviation of inter maxima distances |
| Amplitude mean of maxima/minima |
| Amplitude range of maxima |
| Linear regression slope, offset and quadratic error |
| Quadratic regression a ,b, offset and quadratic error |
| **F0 functionals** |
| Percentage of non-zero frames |
| Mean, max, min, standard deviation of segments length |

*descriptors* (LLDs), which are acoustic features extracted for each frame, such as F0 contour, *Mel-frequency cepstral coefficients* (MFCCs), zero crossing rate and RMS energy. Table 1 lists these LLDs. For each LLD, the toolkit extracts functionals at the sentence level such as mean, maximum and range, creating a 6,373 dimensional feature vector, referred to as *high level descriptors* (HLDs). Table 2 lists the HLDs derived from LLDs. Schuller *et al.* [38] describe this feature set in detail, which we refer to as the ComParE set.

## III. PROPOSED APPROACH

This paper explores the use of neutral reference models to contrast emotional speech. Instead of collecting sentence level statistics, as in Busso *et al.* [22], this study aims to build lexicon-dependent models to compare frame-by-frame acoustic properties of the target speech. This approach aims to uncover local emotion information conveyed in speech. In the ideal case, we would like to contrast an expressive speech with a timely aligned neutral reference signal conveying the same lexical information. Of course, this restrictive approach is not feasible in real applications, since the ideal reference signal is not available. However, advances in speech synthesis provide an opportunity to construct this reference signal that can be directly used to compare the target sentence. This is the precise goal of this paper.

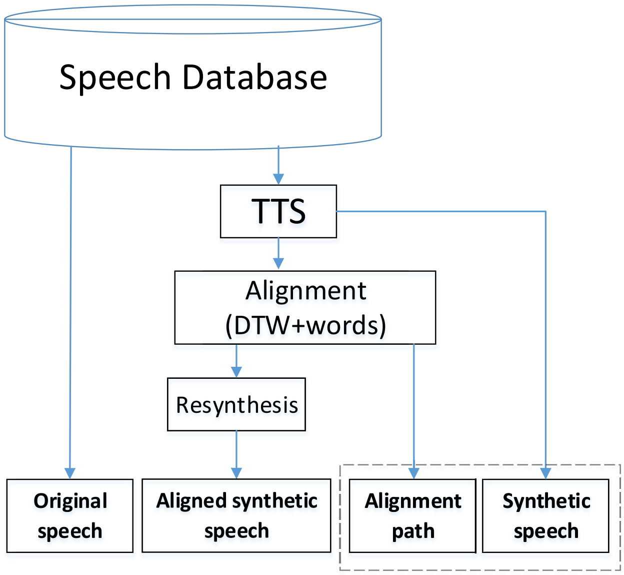


**FIGURE 1.** Overview of the proposed framework to generate a synthetic neutral reference that is timely aligned with the original speech. Section III explains the building blocks.

Figure 1 depicts the overview of the proposed approach, which we briefly summarize before describing the building blocks in detail. It consists of building a reference synthetic speech, which is used to contrast frame-by-frame the target speech (Sec. III-A). This framework is general and can be employed to contrast different acoustic features. As shown in Figure 1, the system takes an input speech from the database with its transcription and word level alignment. The transcription is used to synthesize a speech signal conveying the same lexical information. Multiple variations of the synthesized speech are generated by employing different speech synthesis approaches and also different voices. Having multiple neutral instances of a sentence helps us to suppress the aspects of speech that are not related to the emotional content of speech. It also makes it possible to evaluate the effect of synthesized speech quality on the overall emotion detection performance. Although the generated synthetic speech conveys the same lexical information as the input speech, they are not temporally aligned. Therefore, the synthetic speech is timely aligned to the original natural samples using word boundaries and *dynamic time warping* (DTW).

### A. CREATING SYNTHETIC SIGNALS

As shown in Figure 1, the system takes a target speech with its transcription, in addition to its word alignment. The transcription is used to synthesize a speech signal conveying the same lexical information. This step is implemented with Festival, which is a general multi-lingual speech synthesis system [40]. Instead of building a single synthesized signal for a given target sentence, as in Lotfian and Busso [20], we extract ten realizations by using various TTS methods and different voices. Our goal is to create different versions that better capture the variability found in neutral speech. In particular, this study uses four different TTS methods: HMM-based speech synthesis (HTS), statistical parametric synthesizer using trajectory modeling (CLUSTERGEN), diphone synthesis, and cluster unit selection. We have one voice for HTS, two voices for CLUSTERGEN, two voices for diphone synthesis, and five voices for cluster unit selection.

Notice that the TTS systems are trained with extensive speech samples which are typically emotionally neutral with very few, if any, expressive content. Therefore, we assume that the models are built to generate emotionally neutral speech, and it is expected that the generated synthetic signals provide a good representation of neutral speech (Sec. IV validates this assumption).

### B. TIME ALIGNMENT PROCESS

The main idea of the proposed approach is to compare frame-by-frame low level descriptors derived from the target and synthetic speech signals. The synthesized signals have the same lexical content as the target sentence but they are not timely aligned. Therefore, it is important to estimate the time alignment between both signals. First, the word boundaries of the target and synthetic signals are used to align each of the synthesized signals, keeping the word boundaries of the original speech. The timing of the word boundaries of the target speech is estimated with forced alignment using the transcriptions. The word boundaries of the synthetic signals are provided by the TTS systems. This initial alignment is conducted at the word level, instead of at the phoneme level, since the phoneme set of our acoustic models for forced alignment and the phoneme set used by the four TTS systems are different and a direct mapping cannot be easily established. Furthermore, their dictionaries are also different.

Even after setting the starting time of each word, the alignment is not complete due to word duration differences. Therefore, we rely on DTW to align segments within each word. The allowable region of the dynamic path is set within the range of [1/3,3] [41]. We use the MFCCs as features for DTW, which are estimated for the synthetic signals and the target speech. By combining the word level segmentation and DTW alignment, we build the warping path over each speaking turn.

We use the warping path over each speaking turn to align the synthetic signals. Our approach consists of aligning the speech signals before extracting the features. We use the alignment path as the input of the function overlap-add method [42] implemented in Praat [43], which temporally expands or squeezes the synthetic signals.

Notice that duration is an important prosodic feature to signal emotion (see for example the work of Abdelwahab and Busso [44]). The alignment process will ignore the differences in duration between neutral and emotional speech. To capture this aspect, we estimate the ratio between the speech rate of synthetic and natural speech using the warping path. The relative speech rate is then converted to a logarithmic scale and the resulting curve is smoothed with a 500ms Hamming low pass filter. The relative speech rate is later used as a supplementary LLD for emotion classification. Figure 2 shows an example of a relative speech rate contour.

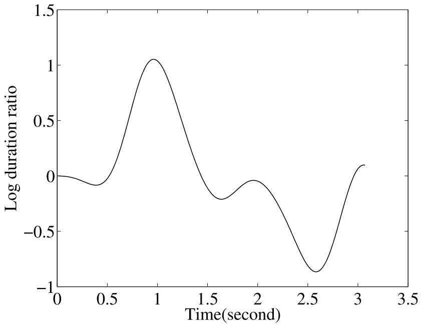


**FIGURE 2. Smoothed speech rate curve for one utterance. The curve gives the localized ratio between the frame durations of the synthetic and target speech signals, expressed in logarithmic scale.**

## IV. ANALYSIS

This section studies the assumption that synthetic speech is a good representation for neutral speech by: a) analyzing the acoustic features from synthetic signals before and after the alignment process (Sec. IV-A), and b) conducting perceptual evaluations to assess emotional content of the synthesized speech (Sec. IV-B). We use the HLDs from the ComParE set.

### A. FEATURE ANALYSIS OF SYNTHESIZED SPEECH

The proposed approach compares acoustic features extracted from the target speech and synthesized signals. Ideally, the selected features should meet the following conditions:

1) Synthetic speech features should be closer to features extracted from neutral natural speech than features extracted from emotional natural speech.
2) The features from the synthetic speech should be robust to the alignment procedure (Sec. III-A).
3) The features from the synthetic speech should maximize the discrimination between neutral (synthetic speech) and emotional (target signal) speech.

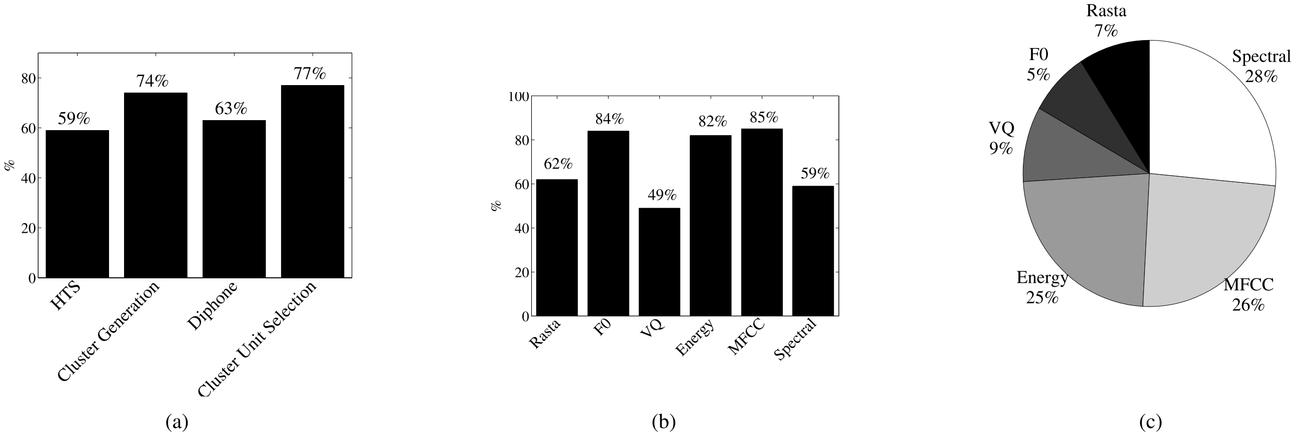


**FIGURE 3. Analysis of criterion one ($r_1$) to assess if the features from the synthesized speech references provide a good representation of natural speech. The figure considers features when $r_1 < 1.1$ (see Eq. 3). (a) Proportion of features that satisfy criterion per TTS method, (b) proportion of features that satisfy criterion per feature group, and (c) distribution per feature group of features that satisfy the condition.**

We analyze the candidate features in terms of these three conditions. Instead of comparing only first or second order statistics of the features, we compare their distributions. For this purpose, we rely on the symmetric version of the *Kullback-Leibler divergence* (KLD) or $J$-divergence [45]. Given two discrete distributions, $p(i)$ and $q(i)$, the $J$-divergence is defined as:

$$J(q,p) = \frac{D(q||p) + D(p||q)}{2} \tag{1}$$

where

$$D(q||p) = \sum_i q(i) \log \frac{q(i)}{p(i)} \tag{2}$$

is the conventional KLD. Since the acoustic features have continuous values, a nonparametric algorithm is used to estimate a discrete distribution for each feature. Nonlinear bins are defined using the K-means clustering algorithm [46], and the values are assigned to the nearest bin center. The bins are not estimated with the SEMAINE database, since emotional differences may bias the location of the bins. Instead, this study uses WSJ corpus, which has neutral sentences. The relatively large size of this corpus provides a robust estimation of the bins. For the analysis, we empirically select 10 bins. There are three different sets: sentences from the SEMAINE corpus, sentences from the WSJ corpus, and synthetic sentences. The three sets consist of multiple speakers, which attenuates the effect of speaker variability. We expect that the variability is mainly due to the TTS process. Using the $J$-divergence measure, we study the feature from these sets in terms of the three conditions.

*Condition 1:* The first condition ensures that features extracted from synthetic speech are not dramatically different from the ones extracted from natural speech (i.e., TTS effectively preserves this feature). The synthetic speech can have artifacts and inconsistencies. Some of these variations has been employed to address the vulnerability of speaker verification systems to synthetic speech, where differences have been observed in prosodic features [47] and phase spectrum [48] (e.g., relative phase shift of different harmonics of voiced speech segments can been utilized to detect synthetic speech due to the loss of phase information during voice conversion [49]). We want to identify features from synthetic signals that are sensitive to these artifacts. This analysis uses sentences from the WSJ corpus as a reference set. We evaluate how natural feature $i$ is using Equation 3:

$$r_1 = \frac{J(synthetic_i, WSJ_i)}{J(SEMAINE_i, WSJ_i)} < 1.1 \tag{3}$$

where $synthetic_i$, $SEMAINE_i$ and $WSJ_i$ are the distributions of the $i^{th}$ feature from the synthetic, SEMAINE and WSJ datasets, respectively. The numerator compares the differences in the distributions of the feature $i$ extracted from the synthetic speech and natural sentences from the WSJ corpus. The denominator compares the differences in the distributions of the features extracted from the SEMAINE and WSJ corpora. This number provides a reference of the expected variation in the feature distributions across natural sentences. The ratio for a ‘‘good'' feature will be less than one, indicating that $J(synthetic_i, WSJ_i) \leq J(SEMAINE_i, WSJ_i)$ (the divergence in feature distribution between synthesized signal and the neutral corpus is less than or equal to the divergence in the feature distribution between natural corpora). Arbitrarily, the features in which the divergence between synthetic and WSJ sentences is 10% higher than the divergence between the SEMAINE and WSJ sentences are considered as unnatural features (see right size of Eq. 3).

Figure 3(a) shows the percentile of features that meet condition one for different speech synthesis approaches. The TTS systems based on cluster generation (74.2%) and cluster unit selection (76.9%) produce synthesized speech with acoustic properties that do not deviate dramatically from natural speech. Overall, over 71.8% of the features satisfy this condition. To understand the acoustic properties that are less affected by the TTS process, we group the features into energy, F0 (fundamental frequency), voice quality, spectral, MFCCs and RASTA features, following the categorization used in Busso and Rahman [50] (while MFCCs and RASTA are spectral features, we decided to keep them in different groups). Figure 3(b) shows the percentile of features per class that meet condition one. Over 80% of the features belonging to the F0, MFCCs, and energy classes meet the criterion. Voice quality features are the acoustic properties that are more affected by the TTS process, where only 49% of the features meet the criterion. Figure 3(c) depicts the contribution of each feature class over the set of selected features across all TTS approaches.

*Condition 2:* The second condition implies that features extracted from the synthetic speech before and after the alignment should remain similar (i.e., alignment process introduced in Section III-A does not affect the feature). We define a second ratio $r_2$, where we estimate the $J$-divergence between features extracted from the WSJ corpus and features from the synthetic speech before ($synthetic_i$) and after ($aligned_i$) the alignment procedure. This condition is illustrated in Equation 4. If a feature is not affected by the alignment process, its distribution after the alignment should remain similar to the distribution extracted before the alignment, and the ratio $r_2$ should be around 1.

$$r_2 = \frac{J(synthetic_i, WSJ_i)}{J(aligned_i, WSJ_i)} \tag{4}$$

We consider that a feature is not affected by the alignment process if $0.9 < |r_2| < 1.1$. Figure 4 shows the proportion of the individual features per feature group that satisfies this condition. The figure shows that voice quality and spectral features are more vulnerable to the alignment process. In contrast, most of the features from F0 and energy groups (i.e., prosody features) satisfies this condition. Overall, the distortion caused by time-scaling the signal only affects 8.6% of the features.

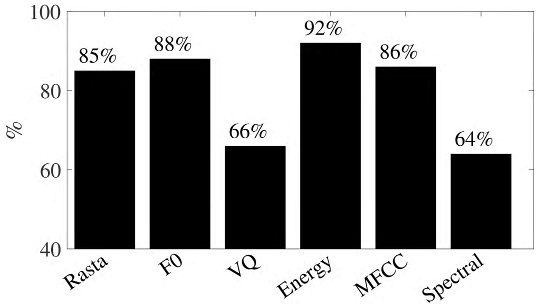


**FIGURE 4. The $J$-divergence between the aligned synthesized speech and the synthesized speech for different feature classes. The figure lists the proportion of features per feature group where $r_2$ is between 0.9 and 1.1 (see Eq. 4).**

*Condition 3:* The third condition implies that the selected features should discriminate between neutral and emotional speech. For a neutral sentence in the corpus, the ideal feature $i$ ($neutral_i$) should have a distribution similar to the one estimated from the aligned synthetic speech, $aligned_i$ (neutral reference). Therefore, the value of $J(aligned_i, neutral_i)$ should be as small as possible. For an emotional sentence, in contrast, the distribution of the feature $i$ ($emotion_i$) should differ from the distribution of the feature derived from the aligned synthetic speech, $aligned_i$. Therefore, the value of $J(aligned_i, emotion_i)$ should be as large as possible. Considering these observations, we define the following ratio:

$$r_3 = \frac{J(aligned_i, emotion_i)}{J(aligned_i, neutral_i)}. \tag{5}$$

High values of $r_3$ will indicate that the feature $i$ is emotionally discriminative. Notice that this ratio is a better indicator than the absolute value of $J(aligned_i, emotion_i)$, which may be sensitive to the mismatch between the original and synthetic signals.

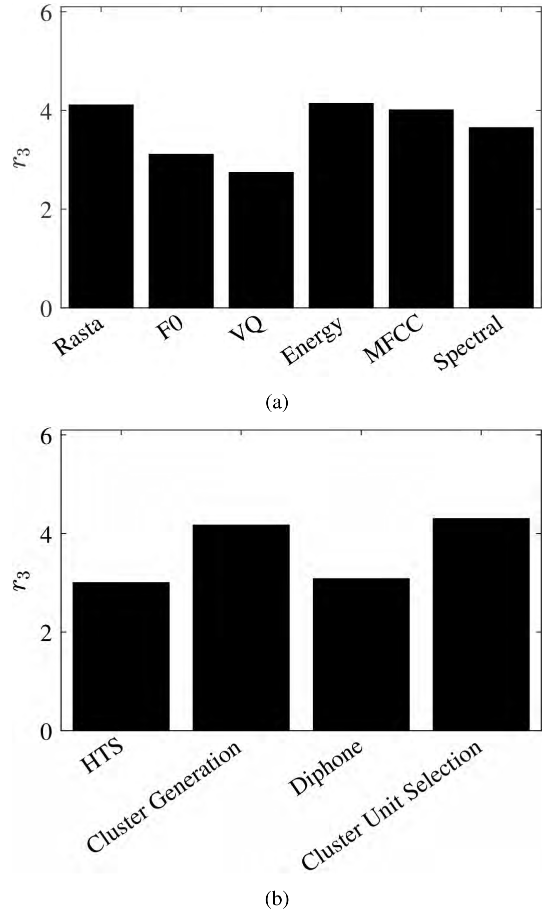


**FIGURE 5. The median emotional discrimination ratio using $r_3$ (see Eq. 5). The figure shows the results in terms of feature groups and TTS methods. (a) Median $r_3$ for each feature group. (b) Median $r_3$ for different TTS methods.**

Figure 5 compares the $r_3$ ratio obtained for the acoustic features. Figure 5(a) compares the median $r_3$ ratios for the feature groups. Energy and F0 features provide the highest discrimination ratios. Variations in energy and F0 are associated with changes in arousal level. Figure 5(b) shows the median ratio $r_3$ for different speech synthesis methods. Cluster unit selection and cluster generation provide the highest discrimination ratios. Figure 6 provides the value of $r_3$ for

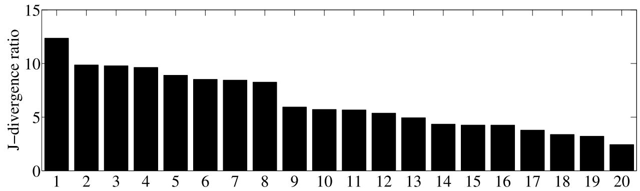


**FIGURE 6.** Top 20 features with the highest discrimination ratio $r_3$ using Equation 5. The names of the corresponding features are listed in Table 3.

**TABLE 3.** Ordered list of the Top 20 features with the highest discrimination ratio $r_3$ shown in Figure 3). Δ denotes the first derivative of the LLDs.

| | Feature (LLD) | Functional (HLD) |
|---|---|---|
| 1 | Spectral energy 1k-4kHz Δ | Relative duration above 90% |
| 2 | Spectral variance Δ | Linear prediction offset |
| 3 | Spectral energy 1k-4kHz Δ | Mean of peak distances |
| 4 | MFCC [1] | Quadradic regression b |
| 5 | MFCC [1] | Quadradic regression a |
| 6 | MFCC [3] | Position of minimum |
| 7 | RASTA auditory spectrum [20] | Standard deviation of segment length |
| 8 | RASTA auditory spectrum [25] | Quadradic regression b |
| 9 | Spectral roll-Off point 0.5 | Percentile range 1% |
| 10 | RASTA auditory spectrum [20] Δ | Second inter-quatile range |
| 11 | MFCC [3] | Maximum segments length |
| 12 | Spectral roll-Off point 0.5 | Percentile range 1% |
| 13 | RASTA auditory spectrum [2] | Minimum segment length |
| 14 | Spectral flux Δ | Position of maximum |
| 15 | RASTA auditory spectrum [6] Δ | Second inter-quatile range |
| 16 | Zero-crossing rate | Minimum segment length |
| 17 | RASTA auditory spectrum [5] Δ | Skewness |
| 18 | RASTA auditory spectrum [5] Δ | Percentile range 99% |
| 19 | MFCC [4] Δ | Minimum segment length |
| 20 | RMS energy | Minimum segment length |

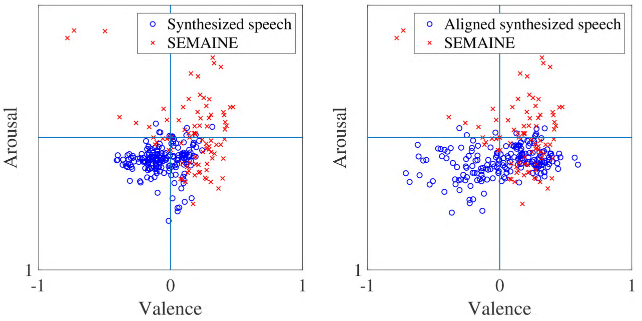
the 20 features with the highest ratio. Table 3 lists the corresponding features including the LLDs and the functionals. Most features in the top of the list correspond to spectral features. The features include five functionals extracted from MFCCs. The list also includes zero crossing rate (feature #16) and RMS energy (feature #20).

### B. PERCEPTUAL EVALUATION OF SYNTHESIZED SPEECH

The feature analysis in Section IV-A shows that some feature extracted from the synthetic speech are similar to the ones extracted from natural speech. This section explores the emotional perception of synthetic speech. The analysis aims to demonstrate that synthetic speech can be used as a neutral reference for emotion recognition.

We annotate the emotional content of synthetic sentences before and after the alignment in terms of valence and arousal scores using subjective evaluations. The scores are compared with the annotations assigned to original sentences. The subjective evaluations are conducted over a subset of the SEMAINE database. We include two sessions for each of the four conditions in the corpus (Obadiah, Spike, Poppy, Prudence), resulting in eight sessions conveying a wide range of emotions (approximately three minutes per session). This set includes 328 sentences from the users. We estimate the emotional content of the synthetic sentences for these eight sessions before and after the alignment.

Each annotator listens to ten sessions. Each evaluator is required to annotate two natural sessions, and eight sessions with synthetic speech (four sessions before the alignment and four sessions after the alignment). We only replace the user's turns for session with synthetic speech. In total, ten listeners participated in this experiment, where 2.5 evaluators annotated natural sessions, and five evaluators annotated the synthetic sessions. The evaluations are designed similar to the subjective evaluations of the SEMAINE database. The annotators are asked to listen to the full session, recording their emotional perception. The evaluations are conducted with the G-trace tool [51], using a joystick. The annotators separately evaluate arousal and valence scores. By default, the position of the joystick tends to return to the center, which is associated with neutral content. This approach reduces the ''inertia'' of staying on an emotional region after the stimulus has passed. The evaluators can hear both the operator (natural speech) and the user (natural speech, synthetic speech, or synthetic speech after the alignment). The natural sessions are used to calibrate the mean and standard deviation across the evaluations. These parameters are used to normalize the emotional traces of the dialogs with synthetic speech, compensating for the bias across evaluators. We calculate the average rating across evaluators over the user turns for each condition, discarding the operator turns.



(a) (b)

**FIGURE 7.** Results of the perceptual evaluation displayed on the arousal-valence space. The figures illustrate the emotional score assigned to the original sentences in the SEMAINE database, and the synthesized sentences created in this study. (a)Synthesized speech. (b) Aligned synthesized speech.

Figure 7 shows the average perception of arousal and valence for the synthetic sentences (Fig. 7(a)) and synthetic sentences after the alignment (Fig. 7(b)). The scores from the original sequences are displayed in red, and are included in Figures 7(a) and 7(b). The figures show that the synthesized samples tend to have slightly negative arousal with valence around zero. The synthetic samples are distributed around the center of the arousal-valence space (94% of the samples are in *region* 5 in Fig. 8). In contrast, many of the sentences from the SEMAINE corpus have more extreme arousal and valence values (only 52% of the samples belong

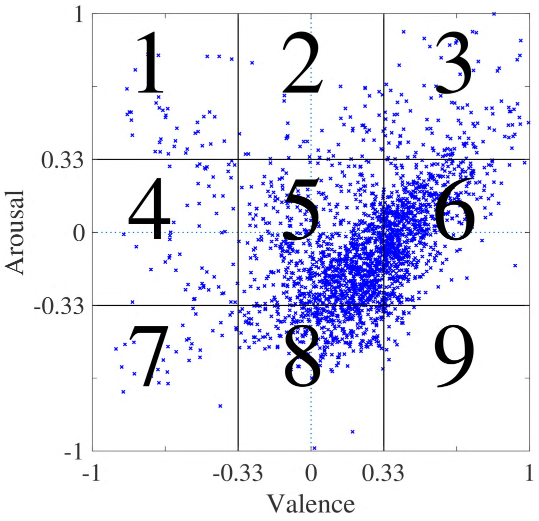


**FIGURE 8.** Distribution of the emotional content of the SEMAINE database. The arousal-valence space is split into nine regions, which are used to formulate different binary classification problems for speech emotion recognition.

to *region* 5 in Fig. 8). Ideally, the evaluators should rate only the acoustic properties. In this case, we would expect that most of the synthesized sentences should be close to the center region. In practice, evaluators use multiple cues including lexical content which is still included in the synthetic sentences. Therefore, even if the acoustic emotional content is neutral, evaluators can still perceive the sentence with some emotion. After the alignment, Figure 7(b) shows that the synthetic sentences are more spread along the valence axis. Speech duration is an important cue to express emotion [44]. Therefore, adding the alignment changes the emotional content of the aligned synthetic speech. Even in this case, most of the sentences are in the neutral region (86% of the samples are in *region* 5 in Fig. 8).

## V. EMOTION CLASSIFICATION FRAMEWORKS

Section IV shows that synthetic speech can serve as a neutral reference signal. For many acoustic features, the approach can be used to increase discrimination between neutral and emotional speech. While this framework can be useful in many domains in affective computing, this study explores its use in speech emotion recognition. This section explains our proposed emotion recognition framework which incorporates neutral reference sentences created with synthetic speech.

Figure 9(a) shows the proposed approaches to incorporate the synthetic references. The method aim to compensate for the lexical content, highlighting the emotional content in the sentences. The approach relies on the whitening transformation.

### A. FEATURE NORMALIZATION WITH THE WHITENING TRANSFORMATION

In order to increase the class separability between emotion classes using synthetic reference, we propose to normalize the features with respect to the ten synthetic reference signals. We propose to use the whitening transformation proposed by Mariooryad and Busso [21], which effectively removes the first and second order statistics of the lexical content.

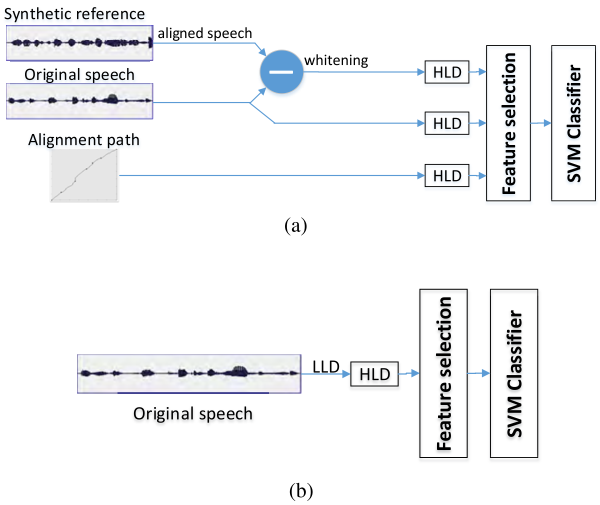


**FIGURE 9.** The figure describes the proposed framework to use the reference signals in speech emotion recognition tasks. The approach uses the aligned synthesized speech to normalize the lexical content of the sentences using the whitening transformation. The baseline method is built with HLDs from the ComParE feature set. (a) Speech emotion classifier using the proposed reference signals. (b) Baseline framework using HLDs (ComParE feature set).

Mariooryad and Busso [21] proposed a framework to factorize idiosyncratic, emotional and lexical factors on speech. The approach created a separate transformation for each phoneme. Consider a LLD $s$ such as the fundamental frequency or MFCCs. The approach creates a trajectory model by parameterizing its shape with a 10D vector ($\mathbf{x}$), obtained after interpolating and resampling the temporal shape of the acoustic feature over the given phoneme. This approach is applied to each phoneme in the sentences for all the emotions and speakers in the corpus. Given the lexical dependent trajectory vectors for a given phoneme, the whitened trajectory is calculated by applying the transformation in Equation 6:

$$\mathbf{x}^w = D_s^{-\frac{1}{2}} V_s'(\mathbf{x} - \mu_s) \tag{6}$$

where $D_s$ and $V_s$ are matrices with the eigenvalues and eigenvectors of the covariance matrix $\Sigma_s$, and $\mu_s$ is the mean vector of the trajectory vectors. This step decorrelates the elements of the vector $\mathbf{x}$, which convey the lexical information associated with the given phoneme.

This whitening transformation is ideal to compensate for the lexical information using the family of synthetic reference signals. The key idea is to estimate the parameters $\Sigma_s$ and $\mu_s$ using the set of aligned synthetic signals created for each sentence. The matrices $D_s$ and $V_s$ are obtained from $\Sigma_s$. This transformation only compensates for consistent patterns across the family of synthetic speech reference signals. As these references are emotionally neutral, the emotional cues are not affected. An important difference in our implementation is the temporal window used to estimate the trajectory vectors. Mariooryad and Busso [21] applied the transformation for each phoneme. Since our reference signals are temporally aligned, we can apply the transformation using a window of any size, as the lexical content is the same as the original sentences. Since the average duration of a phoneme is about 100ms, we fix the window size of the whitening transformation to 100ms. Since the LLDs are estimated every 10ms, we directly obtain a 10D vector with the actual values of the LLDs, avoiding the interpolation and resampling steps.

The dimension of **x** is 10, so the covariance $\Sigma_s$ is a $10 \times 10$ matrix. Since we only use 10 synthetic speech references to estimate $\Sigma_s$, the matrix can be singular or poorly estimated. One potential solution to avoid this problem is to use the ridge regression approach [52], which estimates the covariance matrix as $\Sigma = \frac{1}{N}(\mathbf{x}-\mu)(\mathbf{x}-\mu)' + kI$, with $k \geq 0$ ($N$ is number samples). In our case, we find a global covariance matrix per feature $\Sigma_g$, which is employed as a starting point to estimate $\Sigma_s$:

$$\Sigma_s = \alpha \frac{1}{N}(\mathbf{x}-\mu_\mathbf{s})(\mathbf{x}-\mu_\mathbf{s})' + (1-\alpha)\Sigma_g \tag{7}$$

The global covariance matrix is estimated with $\Sigma_g = \frac{1}{N}(\mathbf{x}_g-\mu)(\mathbf{x}_g-\mu)'$, using 10,000 100ms-windows extracted from sentences from the WSJ corpus. The parameter $\alpha$ is empirically set to 0.9. The resulting whitened low level descriptors are then used to calculate the HLDs at the utterance level.

### B. BASELINE FRAMEWORK

The proposed framework is compared with a classifier trained following a common approach used for emotion classification. Figure 9(b) shows a diagram, where LLDs are extracted from the audio. Then, we extract HLDs creating a 6,373 feature vector (see Sec. II-C). We reduce the feature dimension of the vector using a two layer feature selection approach. The first layer reduces the set using information gain ratio, which reduces the number of feature to 500. The second layer is implemented with the forward-backward feature selection method by maximizing the accuracy of a classifier on the development set. We reduce the feature dimension to 150 for all the experiments. The resulting feature vector is used as the input of an SVM classifier.

## VI. CLASSIFICATION RESULTS

We evaluate the benefits of using the proposed normalization scheme using the reference sentences by measuring the performance of speech emotion binary classifiers built with different feature sets. We measure performance with the F-score, which is calculated using the average precision and average recall rates across both classes. Several binary classification problems have been defined over the valence-arousal space. Figure 8 shows the distribution in the valence-arousal space of the average scores assigned to the speaking turns in the SEMAINE database. The figure shows nine regions, which we use to formulate the binary classification problems.

We consider 20 speakers from the SEMAINE database. We create a development set with data from seven speakers. We exclusively use this set to select a reduced set of features. The data from the remainder 13 speakers is used for the train and test sets using a *leave-one-subject-out* (LOSO) cross-validation approach. In each fold, data from 12 subjects are used for training the models, and data for the remainder speaker are used for testing the results. We report the average results across the 13 folds. For simplicity, the classifiers in the experimental evaluation are implemented with *support vector machine* (SVM) with linear kernel, trained with *sequential minimal optimization* (SMO). The SVMs are implemented in WEKA [53]. The complexity parameter for the binary emotion SVM classifiers is set to $c = 1.0$, following the settings provided by previous studies in speech emotion recognition [54]. The formulations of the emotion recognition problems considered in this study include cases with imbalanced classes. We compensate for the highly imbalanced classes using the *synthetic minority over-sampling technique* (SMOTE) [55], creating balanced classes.

Since the dimension of the ComParE feature set is very large, we reduce the number of features following a two-step feature selection approach for each classification task, using the development set. The first step reduces the original number of features to 1,000 by applying *information gain* (IG). This entropy-based approach independently considers each feature, so it is very efficient. The second step reduces the feature vector to 150 features using a wrapper-based forward-backward approach by maximizing the performance of the SVM classifier on the development set. We consistently follow this approach, creating classifiers trained with 150 features across conditions.

### A. DISCRIMINATIVE ANALYSIS

This section analyzes the performance of binary classifiers that discriminate between high and low values of arousal and valence. For arousal, we consider *regions* (1,2,3) for high arousal and *regions* (7,8,9) for low arousal (see Fig. 8). For valence, we consider *regions* (1,4,7) for low valence and *regions* (3,6,9) for high valence. This approach discards ambiguous samples between classes, attenuating one of the main problems of dichotomizing interval labels into discrete classes [56].

**TABLE 4. Average F-scores for speech emotional classifiers trained with different feature sets. All the classifiers are trained with 150 features after feature selection.**

| Feature set | F-score[%] | |
|---|---|---|
| | Arousal | Valence |
| Baseline | 81.38 | 65.13 |
| Aligned speech (W) | 81.96 | 63.69 |
| Aligned (W) & Speech rate | 83.05 | 64.27 |
| Aligned (W) & Baseline | 81.97 | 65.83 |
| Aligned (W) & Speech rate & Baseline | 84.11 | 66.92 |

Table 4 lists the average F-score of classifiers trained with different features sets. The first row describes the performance of the baseline system trained with the 150 features selected from the ComParE feature set described in Section II-C. On average, the baseline system achieves a 81.38% F-score for arousal and a 65.13% F-score for valence. The second row of Table 4 shows the performance when we use the aligned synthesized speech to contrast emotional speech. This set uses the whitening transformation to compensate for the lexical information in the LLDs, as described in Section V-A. After the normalization, we extract HLDs obtaining a 6,373D feature vector. We reduce the dimension to 150 features using feature selection. This feature set has similar performance to the baseline method for arousal, but slightly lower F-score for valence. The third row in Table 4 shows the performance achieved when the aligned speech feature set is expanded with the speech rate features. The speech rate feature set is generated by applying functionals to the speech rate contour and its first order derivative (Sec. III-B). We rely on the 39 functionals applied to the F0 contour in the ComParE feature set, creating a 78D feature set. Combining the speech rate features with the features from the aligned speech after the whitening transformation improves the average F-score. The improvement is larger for arousal (1.1% absolute gain). When we combine the features from the aligned speech using the whitening transform with the baseline features, the F-scores improves over the baseline, especially for valence. Notice that these classifiers are also trained with 150 features selected from the pool of 12,746 features (i.e., baseline + aligned speech). Interesting, 41% (arousal) and 46% (valence) of the selected features come from the aligned speech feature set, indicating that this feature set is discriminative, and complementary to the baseline set. The last row of Table 4 corresponds to the classifier trained with the aligned speech feature set, the baseline feature set, and the speech rate feature set. After selecting 150 from the pool of features, we obtain the best performances, which improve the F-score of the baseline system by 2.73% (absolute) for arousal, and 1.7% (absolute) for valence. From the 150 features selected from this pool of features, 51% (arousal) and 56% (valence) come from the baseline set, 45% (arousal) and 42% (valence) from the aligned speech speech set, and 4% (arousal) and 2% (valence) from the speech rate feature set. We observe that the aligned speech feature set, and the speech rate features provide complementary information that increases the performance of the system.

### B. ANALYSIS OF THE WHITENING TRANSFORMATION

This section evaluates the proposed whitening transformation to compensate for the lexical content using the family of aligned synthetic reference signals. We evaluate two alternative normalization schemes. The first baseline approach to normalize the lexical content scales the LLDs. This transformation assumes that $\Sigma_s = \sigma_s I$, where $I$ is the identity matrix. Under this assumption, Equation 6 becomes:

$$\mathbf{x}^s = \begin{bmatrix} \frac{1}{\sigma_s} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_s} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{\sigma_s} \end{bmatrix} (\mathbf{x} - \mu_s) = \frac{1}{\sigma_s} I(\mathbf{x} - \mu_s) \tag{8}$$

In this approach, referred to as *scaling*, we find the standard deviation for each feature of the extracted LLD from its ten reference signal ($\sigma_s$). Then, we normalize the features using the transformation in Equation 8.

The second baseline approach to normalize the lexical content subtracts the LLDs. One straightforward approach to normalize the speech signal using the aligned synthesized reference is to calculate the difference between the features extracted from both signals. This approach was used in Lotfian and Busso [20], which used only one synthetic reference signal. We use a variation of this method to compare the performance of the proposed whitening transformation. Since this study uses ten synthetic signals as references, we estimate the difference between the LLDs extracted from the original speech and the average LLDs extracted from its ten reference signals. The HLDs are then extracted from the features after the subtraction. This subtraction based normalization approach removes the average content due to lexical variability, but it does not capture higher order statistics as the proposed whitening transformation.

The evaluation in this section also considers the binary classification problems defined in Section VI-A. Figure 10 reports the results using the *whitening*, *scaling* and *subtraction* approaches using different feature sets. When compared to the baseline approaches to normalize the lexical content, the figure indicates that the whitening transformation provides the best results for arousal and valence, where the differences are statistically significant, as indicated by the asterisks above the bars (one-tailed t-test, $p$-value $\leq 0.05$). We use the whitening transformation for the rest of the experiments.

### C. EMOTION RECOGNITION FORMULATIONS

We also evaluate the proposed approach on several binary classification tasks defined over the arousal and valence space. The purpose of this analysis is to analyze the emotional content that our approach is able to effectively contrast. In addition to the binary classification problems described in Section VI-A, we consider *region* 5, which includes sentences in the center of the arousal-valence coordinate (i.e., neutral speech), versus each of the other eight regions (see Fig. 8). We only implement the task if we have at least 50 sentences in each class. We discard *regions* 7 and 9 due to this requirement (i.e., six binary classification tasks).

Table 5 shows the average F-score for the binary classification tasks defined using the regions in Figure 8. The table shows that features extracted from the aligned synthesized speech are effective to improve the performance over the baseline feature set. We consistently observe this result for all the classification tasks considered in this study. Table 5 also shows higher performance in classification tasks between two regions with different arousal (e.g., *regions* 5 versus 2). The F-scores are reduced when regions have similar arousal scores, but different valence scores (i.e., *regions* 5 versus 4, or *regions* 5 versus 6). For these tasks, the F-score is about 10% (absolute) lower. These results suggest that our synthetic speech reference is more effective in contrasting

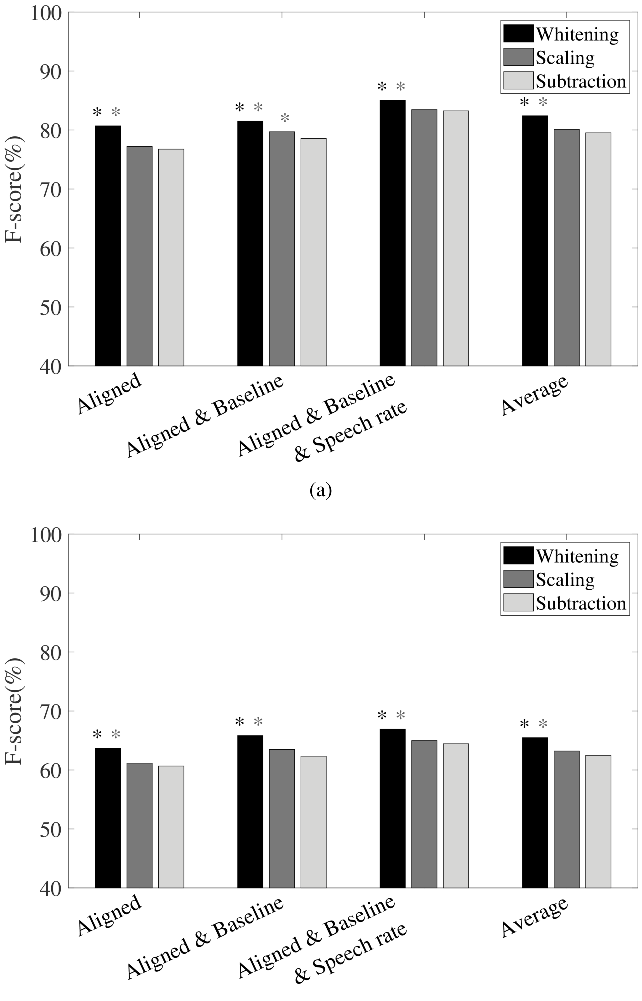


**FIGURE 10.** Comparison between the whitening transformation with the two alternative methods introduced in Section VI-B (scaling and subtracting methods). An asterisk on top of a bar indicates that one approach outperforms the method indicated by the color of the asterisk, asserting significance at *p*-value≤0.05. (a) Arousal. (b) Valence.

**TABLE 5.** F-score of binary classification problems formulated by considering different regions in the arousal-valence space (Fig. 8). For each task, the binary classes are balanced using SMOTE (B: *baseline*, A: *aligned speech feature set*, SR: *speech rate*).

| | | | F-score[%] | | | |
|---|---|---|---|---|---|---|
| Class 1 | Class 2 | Class size | B | A | B& A | B & A & SR |
| 1,2,3 | 7,8,9 | 91 /454 | 81.38 | 81.96 | 82.51 | **84.93** |
| 1,4,7 | 3,6,9 | 107/516 | 65.13 | 63.69 | 65.93 | **66.54** |
| 5 | 1 | 1640/53 | 69.42 | 68.50 | 71.35 | **71.56** |
| 5 | 2 | 1640/73 | 68.58 | 64.28 | 69.13 | **69.57** |
| 5 | 3 | 1640/174 | 69.67 | 68.84 | 71.58 | **72.81** |
| 5 | 4 | 1640/62 | 62.69 | 61.42 | 63.48 | **63.62** |
| 5 | 6 | 1640/342 | 59.13 | 59.28 | **60.43** | 60.31 |
| 5 | 7 | 1640/38 | — | — | — | — |
| 5 | 8 | 1640/142 | 69.87 | 69.11 | 70.16 | **70.29** |
| 5 | 9 | 1640/25 | — | — | — | — |

emotional content that deviates in terms of arousal. These results agree with the perceptual evaluation in Section IV-B, which shows higher variability along the valence domain for the aligned synthesized speech. Notice that finding acoustic features that are discriminative in the valence domain is a challenge task [50], [57]. Despite the higher performance improvement in classification tasks along the arousal domain, employing the synthetic reference is still useful in discriminating between different levels of valence.

### D. ANALYSIS OF REDUCED FEATURE SET

The analysis in Section IV-A identifies emotionally salient features using the ratio $r_3$ (Equation 5). This section investigates whether features with the highest ratio retain the discriminative power in emotion classification problems. This analysis considers the binary problems considered in Section VI-A (i.e., low and high values of either arousal or valence). We consider a reduced subset of the aligned speech feature set that satisfies the condition $r_3 > 1$ (e.g., $J(aligned_i, emotion_i) > J(aligned_i, neutral_i)$). With this criterion, we discard 36.4% of the features. We reduce the set to 150 features per condition using feature selection.

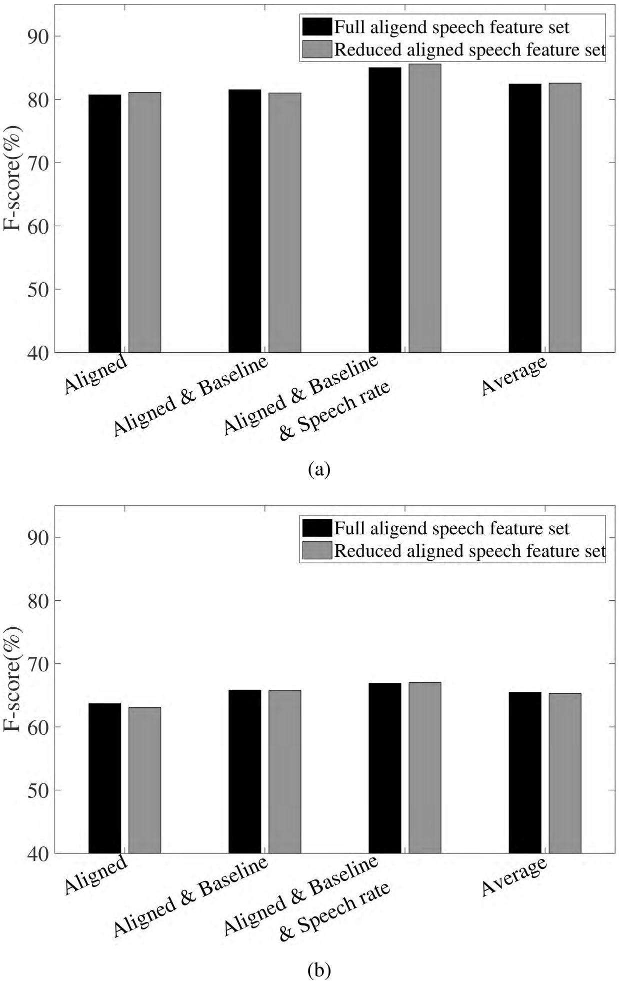


**FIGURE 11.** Comparison of the performance of classifiers trained with the full or reduced aligned speech feature sets. The reduced aligned speech feature set includes only features where $r_3 > 1$ (Eq. 5). (a) Arousal. (b) Valence.

Figure 11 shows the average F-scores of classifiers trained with different feature sets. The F-scores are equivalent when we use the full and reduced aligned speech feature set. The differences in performance are not statistically significant. This result indicates that the criterion three using $r_3$ was effective in quantifying the discriminant information in the features after alignment.

## VII. CONCLUSIONS

The paper proposed a novel framework to create neutral reference models from synthetic speech to contrast the emotional content of a speech signal. The approach creates timely aligned synthetic speech references that convey the same lexical content as the original speech. Since they are aligned, they can be used to contrast frame-by-frame the emotional cues, effectively removing the lexical content of the sentence. We implemented this approach by creating 10 synthetic references for each speech sentence using different TTS approaches. These synthetic sentences preserve many of the acoustic properties of neutral speech and can be used to contrast emotional cues, as demonstrated by the analysis. The perceptual evaluation showed that the synthetic sentences are also perceived with arousal and valence scores similar to the ones assigned to neutral sentences.

To demonstrate one of the potential use of building synthetic speech references in affective computing, we conducted emotion classification evaluations where the family of synthetic speech references were used to remove the lexical content. We considered a feature normalization approach based on the whitening transformation. The results showed absolute improvements of 2.73% (arousal) and 1.7% (valence) in the average F-score, when the features extracted from the aligned speech were added to the feature set. The complementary information provided by the proposed features increases the performance of speech emotion classifiers.

The proposed approach assumes that the lexical information in the sentence is known. This assumption holds for non-real time scenarios in which the transcriptions are available (e.g., analysis of jury trial). In other cases, the lexical information has to be inferred from speech by using *automatic speech recognition* (ASR). Our future work includes the study of the impact of *word error rate* (WER) in the proposed approach. We expect that the impressive performance achieved by current ASR systems [58], [59] will provide the infrastructure to incorporate the proposed system.

Algorithms that are able to identify localized emotional segments have the potential to shift current approaches used in the area of affective computing. These advances represent a transformative breakthrough in the area of behavioral analysis and affective computing. The findings in this study go beyond improvements in classification performance, demonstrating the feasibility of using advances in speech synthesis to build robust neutral reference models to contrast and study frame-by-frame emotional speech. Having established the base infrastructure for the proposed research, several new scientific avenues will emerge that serve as truly innovative advancements, creating mechanisms to understand better the production and perception of emotions. For example, having the synthetic speech reference can be used to analyze the externalization of emotions. We have shown that emotion is not uniformly distributed across time [4], [5], [31]. This framework can be used to identify localized regions that deviate from neutral behaviors.

## REFERENCES

[1] R. W. Picard, "Affective computing," MIT Media Lab. Perceptual Comput. Sect., Cambridge, MA,USA, Tech. Rep. 321, Nov. 1995.

[2] J. Arias, C. Busso, and N. B. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 278–294, Jan. 2014.

[3] J. P. Aris, C. Busso, and N. B. Yoma, "Energy and F0 contour modeling with functional data analysis for emotional speech detection," in *Proc. Interspeech*. Lyon, France, Aug. 2013, pp. 2871–2875.

[4] C. Lee *et al.*, "Emotion recognition based on phoneme classes," in *Proc. 8th Int. Conf. Spoken Lang. Process.* Jeju Island, Korea, Oct. 2004, pp. 889–892.

[5] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *Proc. Interspeech Eurospeech*, Antwerp, Belgium, Aug. 2007, pp. 2225–2228.

[6] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social Emotions in Nature and Artifact: Emotions in Human and Human-Computer Interaction*, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford Univ. Press, Nov. 2013, pp. 110–127.

[7] E. Mower *et al.*, "Interpreting ambiguous emotional expressions," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.* Amsterdam, The Netherlands, Sep. 2009, pp. 1–8.

[8] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 12, pp. 2423–2435, Dec. 2018.

[9] R. Cauldwell, "Where did the anger go? The role of context in interpreting emotion in speech," in *Proc. ISCA Tutorial Res. Workshop (ITRW) Speech Emotion*. Newcastle, Northern Ireland, U.K., Sep. 2000, pp. 127–131.

[10] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Commun.*, vol. 40, nos. 1–2, pp. 5–32, Apr. 2003.

[11] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.* San Antonio, TX, USA, Oct. 2017, pp. 415–420.

[12] G. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact.* San Antonio, TX, USA, Oct. 2017, pp. 248–255.

[13] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Trans. Affect. Comput.*, to be published.

[14] H. Martinez, G. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE Trans. Affect. Comput.*, vol. 5, no. 2, pp. 314–326, Jul./Sep. 2014.

[15] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.* Shanghai, China, Mar. 2016, pp. 5205–5209.

[16] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 186–202, Jan. 2015.

[17] R. Lotfian and C. Busso, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 490–494.

[18] S. Parthasarathy, R. Lotfian, and C. Busso, "Ranking emotional attributes with deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 4995–4999.

[19] S. Parthasarathy and C. Busso, "Preference-learning with qualitative agreement for sentence level emotional annotations," in *Proc. Interspeech*, Hyderabad, India, Sep. 2018, pp. 252–256.

[20] R. Lotfian and C. Busso, "Emotion recognition using synthetic speech as neutral reference," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Brisbane, QLD, Australia, Apr. 2015, pp. 4759–4763.

[21] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Commun.*, vol. 57, pp. 1–12, Feb. 2014.

[22] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 582–596, May 2009.

[23] J. Pittam and K. Scherer, "Vocal expression and communication of emotion," in *Handbook Emotions*, M. Lewis and J. Haviland, Eds. New York, NY, USA: Guilford Press, 2008, pp. 185–198.

[24] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[25] B. Schuller *et al.*, "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals," in *Proc. Interspeech Eurospeech*. Antwerp, Belgium, Aug. 2007, pp. 2253–2256.

[26] A. Batliner, D. Seppi, S. Steidl, and B. Schuller, "Segmenting into adequate units for automatic recognition of emotion-related episodes: A speech-based approach," *Adv. Hum.-Comput. Interact.*, vol. 30, pp. 1–15, Jan. 2010.

[27] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schr "oder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *Proc. ISCA Tutorial Res. Workshop (ITRW) Speech Emotion*. Newcastle, Northern Ireland, U.K.: ISCA, Sep. 2000, pp. 19–24.

[28] C. Whissel, "The dictionary of affect in language," *Measurement Emotions*, vol. 4, R. Plutchik and H. Kellerman, Eds. New York, NY, USA: Academic, 1989.

[29] H. Wang, A. Li, and Q. Fang, "F0 contour of prosodic word in happy speech of Mandarin," in *Affective Computing and Intelligent Interaction* (Lecture Notes in Computer Science), vol. 3784, J. Tao, T. Tan, and R. Picard, Eds. Berlin, Germany: Springer, Oct. 2005, pp. 433–440.

[30] M. Goudbeek, J. Goldman, and K. R. Scherer, "Emotion dimensions and formant position," in *Proc. Interspeech*, Brighton, U.K., Sep. 2009, pp. 1575–1578.

[31] C. Busso and S. S. Narayanan, "Joint analysis of the emotional fingerprint in the face and speech: A single subject study," in *Proc. IEEE 9th Workshop Multimedia Signal Process.* Crete, Greece, Oct. 2007, pp. 43–47.

[32] B. Schuller and F. Burkhardt, "Learning with synthesized speech for automatic emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*. Dallas, TX, USA, Mar. 2010, pp. 5150–5153.

[33] B. Schuller, Z. Zhang, F. Weninger, and F. Burkhardt, "Synthesized speech for model training in cross-corpus recognition of human emotion," *Int. J. Speech Technol.*, vol. 15, no. 3, pp. 313–323, Sep. 2012.

[34] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schróder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, Jan./Mar. 2012.

[35] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Trans. Affect. Comput.*, vol. 6, no. 2, pp. 97–108, Apr./Jun. 2015.

[36] S. Mariooryad, and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *Proc. Humaine Assoc.Conf. Affect. Comput. Intell. Interact.* Geneva, Switzerland, Sep. 2013, pp. 85–90.

[37] D. Paul and J. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. 2th Int. Conf. Spoken Lang. Process.* Alberta, Canada, Oct. 1992, pp. 899–902.

[38] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*. Lyon, France, Aug. 2013, pp. 148–152.

[39] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The munich versatile and fast open-source audio feature extracto," in *Proc. ACM Int. Conf. Multimedia*. Florence, Italy, Oct. 2010, pp. 1459–1462.

[40] P. Taylor, A. W. Black, and R. Caley, "The architecture of the festival speech synthesis," in *Proc. 3rd ESCA/COCOSDA Workshop Speech Synth.* Blue Mountains, NSW, Australia, Nov. 1998, pp. 147–151.

[41] S. Ohno, M. Fukumiya, and H. Fujisaki, "Quantitative analysis of the local speech rate and its application to speech synthesis," in *Proc. 4th Int. Conf. Spoken Lang. Process.*, vol. 4, Oct. 1996, pp. 2254–2257.

[42] W. Verhelst, "Overlap-add methods for time-scaling of speech," *Speech Commun.*, vol. 30, no. 4, pp. 207–221, Apr. 2000.

[43] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," Inst. Phonetic Sci. Univ. Amsterdam, Amsterdam, The Netherlands, Tech. Rep. 132, 1996.

[44] M. Abdelwahab and C. Busso, "Evaluation of syllable rate estimation in expressive speech and its contribution to emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*. South Lake Tahoe, CA, USA, Dec. 2014, pp. 472–477.

[45] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[46] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York, NY, USA: Wiley, 2000.

[47] A. Ogihara, U. Hitoshi, and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *Trans. Fundam. Electron., Commun. Comput. Sci.*, vols. E88–A, no. 1, pp. 280–286, Jan. 2005.

[48] Z. Wu, E. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. Interspeech*. Portland, OR, USA, Sep. 2012, pp. 1700–1703.

[49] P. L. D. Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Prague, Czech Republic, May 2011, pp. 4844–4847.

[50] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Proc. Interspeech*. Portland, OR, USA, Sep. 2012, pp. 1179–1182.

[51] R. Cowie, M. Sawey, C. Doherty, J. Jaimovich, C. Fyans, and P. Stapleton, "Gtrace: General trace program compatible with EmotionML," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.* Geneva, Switzerland, Sep. 2013, pp. 709–710.

[52] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[53] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.

[54] A. Hassan and R. Damper, "Multi-class and hierarchical SVMs for emotion recognition," in *Proc. Interspeech*. Makuhari, Japan, Sep. 2010, pp. 2354–2357.

[55] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[56] S. Mariooryad and C. Busso, "The cost of dichotomizing continuous labels for binary classification problems: Deriving a bayesian-optimal classifier," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 119–130, Jan./Mar. 2017.

[57] K. Sridhar, S. Parthasarathy, and C. Busso, "Role of regularization in the prediction of valence from speech," in *Proc. Interspeech*, Sep. 2018, pp. 941–945.

[58] G. Saon *et al.* (Mar. 2017). "English conversational telephone speech recognition by humans and machines." [Online]. Available: https://arxiv.org/abs/1703.02136

[59] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

**REZA LOTFIAN** (S'17) received the B.S. degree (Hons.) in electrical engineering from the Department of Electrical Engineering, Amirkabir University, Tehran, Iran, in 2006, the M.S. degree in electrical engineering from the Sharif University of Technology, Tehran, in 2010, and the Ph.D. degree in electrical engineering from The University of Texas at Dallas, in 2018. He is currently a Research Scientist with Cogito Corporation, Boston, MA, USA. His research interests include the area of speech signal processing, affective computing, human–machine interaction, and machine learning.

**CARLOS BUSSO** (S'02–M'09–SM'13) received the B.S. and M.S. degrees (Hons.) in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is currently an Associate Professor with the Electrical Engineering Department, The University of Texas at Dallas, where he leads the Multimodal Signal Processing Laboratory. He has co-authored the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 Emotion Challenge. His research interest includes human-centered multimodal machine intelligence and applications. His current research interests include the broad areas of affective computing, multimodal human–machine interfaces, nonverbal behaviors for conversational agents, in-vehicle active safety systems, and machine learning methods for multimodal processing. His research has direct implication in many practical domains, including national security, health care, entertainment, transportation systems, and education. He is a member of ISCA, AAAC, and ACM. He was selected by the School of Engineering of Chile as the Best Electrical Engineer graduated, in 2003, across Chilean universities. He received the Provost Doctoral Fellowship, from 2003 to 2005, and the Fellowship in Digital Scholarship, from 2007 to 2008. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He was a recipient of the NSF CAREER Award and the ICMI Ten-Year Technical Impact Award, in 2014. In 2015, his student received the Third Prize IEEE ITSS Best Dissertation Award. He was the General Chair of ACII 2017.